Towards Intelligent Agents to Assist in Modular Construction: Evaluation of Datasets Generated in Virtual Environments for AI training

Keundeok Park, Semiha Ergan, Chen Feng

Tandon School of Engineering, New York University, the USA
E-mail: kp2393@nyu.edu, semiha@nyu.edu, cfeng@nyu.edu

Abstract –

Modular construction aims at overcoming challenges faced by the traditional construction process such as the shortage of skilled workers, fast-track project requirements, and cost associated with on-site productivity losses and recurrent rework. Since manufacturing is done off-site in controlled factory settings, modular construction is associated with increased productivity and better quality control. However, because every construction project is unique and results in distinct work pieces and building elements to be assembled, modular construction factories necessitate better mechanisms to assist workers during the assembly process in order to minimize errors in selecting the pieces to be assembled and idle times while figuring out the next step in an assembly sequence. Machine intelligence provides opportunities for such assistance; however, a challenge is to rapidly generate large datasets with rich contextual data to train such intelligent agents. This work overviews a mechanism to generate such datasets in virtual environments and evaluates the performance of AI models trained using data generated in virtual environments in recognizing the next installation step in modular assembly sequences. Performance of the trained MV-CNN models (with accuracy of 0.97) shows that virtual environments can potentially be used to generate the required datasets for AI without the costly, time-consuming, and labor-intensive investments needed upfront for capturing real-world data.

Keywords –

Scene understanding; Virtual Environment; MV-CNN; Computer Vision

1 Introduction

While productivity in other industries has doubled in the past decades, productivity in the construction industry has remained flat [26-27]. In addition, more challenges are faced with the shortage of skilled workers and tighter construction sites in urban settings that impact productivity in general. As a solution to this situation, the industry shows an interest towards modular construction, whose principle is to preassemble work pieces into volumetric units (or preassembled panels) off-site and stack them on-site. The advantage of off-site production is the controls over the working environment, so it is not affected by weather and site-specific conditions, and there is an opportunity to increase productivity because a relatively fixed production line can be maintained as compared to the conventional construction sequence. In addition, it is environmentally friendly with 15% less construction waste as compared to on-site construction [1]. Despite these advantages, the inherent unique nature of each construction project brings the same challenge to the modular construction processes, too. To enhance the productivity in modular construction factories and minimize rework and idle times of workers while identifying ever-changing work pieces and assembly sequences during manufacturing, intelligent agents can be used for assistance to identify the next step in an assembly sequence for workers. However, training such intelligent agents requires extensive information about features of a construction site, construction processes, and also geometries of assembled pieces [2]. This requirement, coupled with the variations in these due to the uniqueness (e.g., design, materials) of construction projects, increases the cost of gathering a large scale but contextually rich dataset needed for training of reliable AI models. Since quality and quantity of data are essential for training intelligent agents [3] and real-world data (which naturally has all the context needed for the training process) is expensive to capture, there is a need for alternative ways to rapidly generate realistic and context-rich datasets.

In this study, we formulated an approach that leverages virtual environments reconstructed from real factories to rapidly and systematically generate large scale context-rich datasets and provided the results of a predictive model built for recognizing the assembly step in a sequence using the dataset generated in virtual environments.

2 Literature review

This section provides (a) an overview of datasets generated by leveraging virtual environments for AI training, and (b) a brief synthesis of research studies at the intersection of AI and modular construction.

2.1 Overview of datasets generated by leveraging virtual environments

Although real world generated datasets effortlessly represent the rich complexity and the context required for AI to learn properly, they are expensive and require a longer upfront time to capture them. Virtual environments on the other hand provide a huge advantage of rapidly replicating real environments and representing a wide variety of objects in scenes. Realistic representation of the real environment in the virtual space is critical for data quality obtained from these environments. There are generally two main sources to bring reality to virtual environments: (1) a scanned environment [6-8,19,22], where a real environment is scanned and reconstructed as-is; and (2) a synthetic environment [5,20-21], where virtual environments are constructed through 3D modelling from scratch that reflect the principles of real world such as lighting, textures, and colours.

Various datasets have been generated in virtual environments for scene understanding and object recognition purposes. Since the first generation of large dataset are limited to the 2D perspective images, many research groups put efforts on establishing large-scale 3D datasets that contain rich contextual information about the scenes to provide the datasets and benchmark systems to improve the scene understanding. These next generation 3D datasets are captured in different settings; such as scenes from urban environments [18], indoor settings from households [5], and indoor settings from offices [7] and can be applied to various problems such as object detection, scene understanding, and room layout estimation. Also, since the type of intelligence (e.g., self-driving, construction robot, household AI) expected from AI is determined based on the context of a dataset, the 3D environment should provide rich contextual information to AI. For example, for a construction robot to be tasked with understanding the next step in an assembly sequence, the dataset should include variations in geometric shapes that assemblies contain, variations in the views the sequence is captured, variations in the background where modules are assembled such as the location, surrounding objects, and congestion.

For a smooth learning workflow, data should be well-structured with RGB-D (Red, Green, Blue and Depth), annotations on objects in scenes (e.g., 2D-3D bounding boxes and classification labels, semantic labels), relationships between objects in scenes (e.g., scene graph generation from objects) [4]. Actively used datasets for experiments include, Matterport3D [6], AI2-THOR [5], Gibson env. [7], and Replica Dataset [8]. These datasets can be widely used for simulation such as scene understanding, robot navigation, and robotic manipulation. For instance, Gibson env. was utilized for creating 3D scene graphs containing semantic information of household furniture and rooms to serve the purpose of scene understanding [4]. Matterport3D provides 194,400 RGB-D real-world captured indoor images primarily for scene understanding purposes. On the other hand, AI2-THOR is providing digital models of 89 apartments with 600 objects and interactions such as opening refrigerator. The purpose of this dataset is to train robot manipulator in the household context. Subsequently, Gibson env. provides laser scanned data of 572 buildings which has semantics, depth, and normal of faces. This data aims to train sensorimotor robot AI models. Likewise, the Replica contains 18 scanned apartments models with depth and semantics of objects for scene understanding.

The commonality of these datasets is that they were generated in virtual space that resembles the real-world for giving sufficient sensory information to AI to have perception to solve problems such as scene understanding, robot navigation, and robotic manipulation. In this paper, we reconstruct the 3D modular factory environment based on the actual factory and train AI model to understand the work progress in the modular construction factory context.

2.2 Overview of AI based research in the modular construction domain

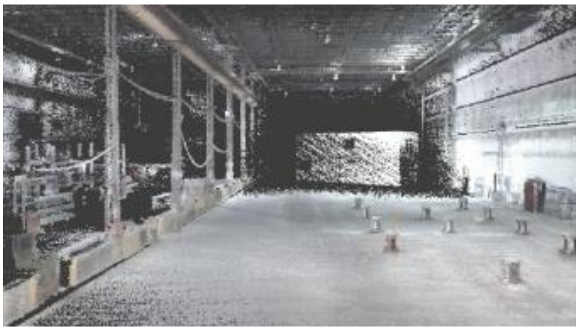
Related research studies reveal that, machine learning techniques have been applied to various challenges such as automatic identification of construction activities [16,23-24], classification of subtypes of BIM objects [9,25], and detection of modules that are being stacked at a construction site [17]. In relation to modular construction domain, several machine learning algorithms such as Support Vector Machine (SVM) on video and audio data [16], LSTM [23-24], and Multi-View Convolutional Neural Networks (MV-CNN) [9] have been evaluated. High model performances in these studies showed the applicability of deep learning in construction related problems with a relatively small dataset and computing resources by transfer learning from the models trained with larger datasets from the computer vision domain with an F-1 score of up to 0.93. However, the models are limited to inference of generic objects (e.g., door, wall) and the usability of these models to the problem of assembly sequence identification and classification is yet to be evaluated. However, in general previous research studies provide a sound point of departure, indicating (a) an opportunity to utilize feature engineering to enhance the classification results, (b) a high performance of models that utilize 2D perspective images captured from 3D models, and (c) a high performance of models in real world settings that are trained with synthetic data originated from virtual environments.

3 Training AI for recognizing steps in module assembly sequences

In this section, we overview our approach to generate synthetic data and then train vision-based AI progress classifier within virtual factory environments. We generated data from a virtual factory environment and trained a classification model to classify an recognize steps in module assembly sequences. Module assembly progress classifier detects the assembly step of a volumetric unit (or a panel assembly) given any stage of the assembly. So, this intelligent agent will detect the assembly step when the agent navigates in a factory setting and sees a module that is being assembled. This is an essential step in robotic assistance during the assembly process for determining the next piece that is needed in the sequence of an assembly.

3.1 Data Generation Platform

We leveraged VR environments to generate large amounts of data with minimum reality-gap and in a short time frame. Generating realistic virtual environments (e.g., scanning, reconstructing, etc) has an upfront time investment that is justified by the a) elimination of relocation of the camera system from station to station in a factory setting or having to purchase a large collection of cameras to cover all stations; b) elimination of the time wasted while waiting for the real fabrication timing of modular pieces in order to completely capture assembly sequences; and c) elimination of bounding to the fixed number of viewpoints, as experienced while capturing real images in factories.



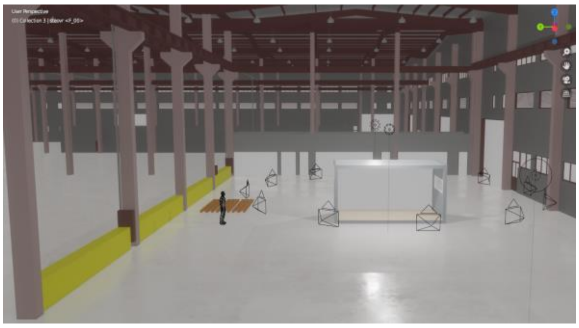


Figure 1. Scanned modular construction factory (top); Virtual factory environment (bottom)

In order to generate datasets in virtual settings, we first created 3D factory environments from a modular construction factory that was scanned with terrestrial laser scanners and converted them into VR (Figure 1). This environment has been detailed in [12].

For simulating the module assembly processes, we obtained real volumetric module designs and separated the volumetric units into panels and work pieces in BIM authoring tools, and then exported them as IFC files. The volumetric modules in IFC format were integrated into the virtual factory environment while the virtual factory environment was reconstructed. Within the virtual environment, multi viewpoint images containing labels (as the steps of the module assembly sequence) were generated using scripts implemented in a 3D graphics tool's API. Subsequently, generated images were split into training and test dataset using the 8:2 ratio.

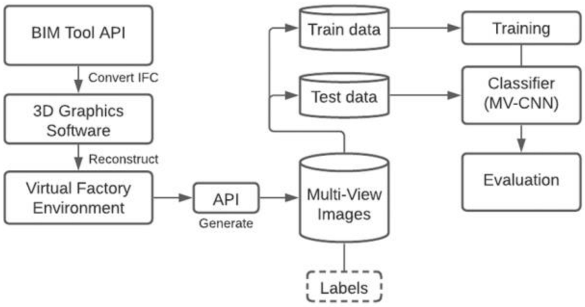


Figure 2. Overview of data generation and training process

To capture multi-view images in the virtual environment, we set virtual cameras that focused on the volumetric modules (see Figure 3). The details of the entire multi-view camera-based platform that was developed by the authors are provided in [12]. In a nutshell, this platform is composed of a randomized camera system and rendering tools. The data is generated by systematically rotating camera views from 12 distinct locations. Using this platform, we generated 84,000 images and separated them into training and testing data set using the 8:2 ratio.

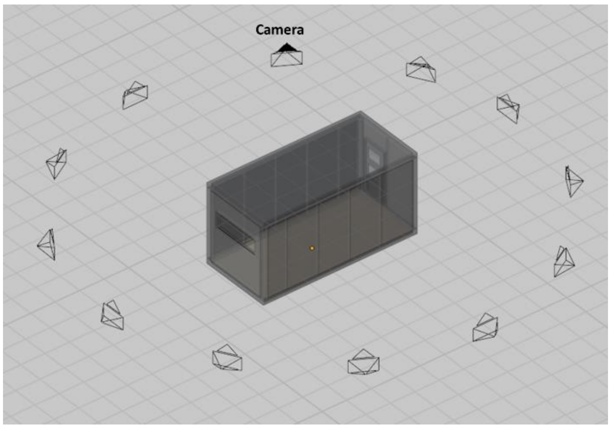


Figure 3. 12 viewpoint camera system to capture images

3.2 Multi-View Convolutional Neural Networks (MV-CNN)

There are two main approaches to provide 3D geometric information to deep learning to identify given objects: 1) point clouds-based approach, which directly inputs raw point clouds [15]; 2) view-based approach, which inputs images from multiple viewpoints [10]. There is essentially no difference between these methods for object classification problems [8,13-14] even though point clouds have more accurate geometric information in 3D coordinates with the disadvantage of time consuming and expensive data capturing process. Furthermore, earlier studies showed that, MV-CNN has a higher overall accuracy as compared to point clouds-based model or machine learning model (SVM) to identify and label 3D model elements [8,13-14]. Hence, in this study, MV-CNN was adopted to build the model to identify and label the steps in a module assembly sequence.

In a nutshell, MV-CNN enables 3D shape recognition by retrieving geometric features from multiple 2D images and combining them into a single set of features. Each image is processed through CNN_1 and pooled over images from multiple viewpoints process through CNN_2 for shape descriptor [10]. A single input for MV-CNN architecture in this work consists of 12 images captured from multiple viewpoints at a location and the corresponding label for those 12 images, which indicates the module assembly step (see Figure 4) (i.e., 1 input). We retrained the model upon the pre-trained MV-CNN that utilized generic 3D objects for training (e.g., chair, sofa, door, etc.).

3.3 Ensuring context variation: sequence of module assembly, background environment

The sequence of volumetric unit assembly used in this study is shown in Figure 5. The volumetric unit is composed of 14 pieces (e.g., frame chassis, wall panel, wall panel with door). Therefore, the generated dataset includes volumetric module assembly sequence captured at different times during the assembly process, at different backgrounds and from viewpoints while these fourteen 3D objects were being assembled in the factory.

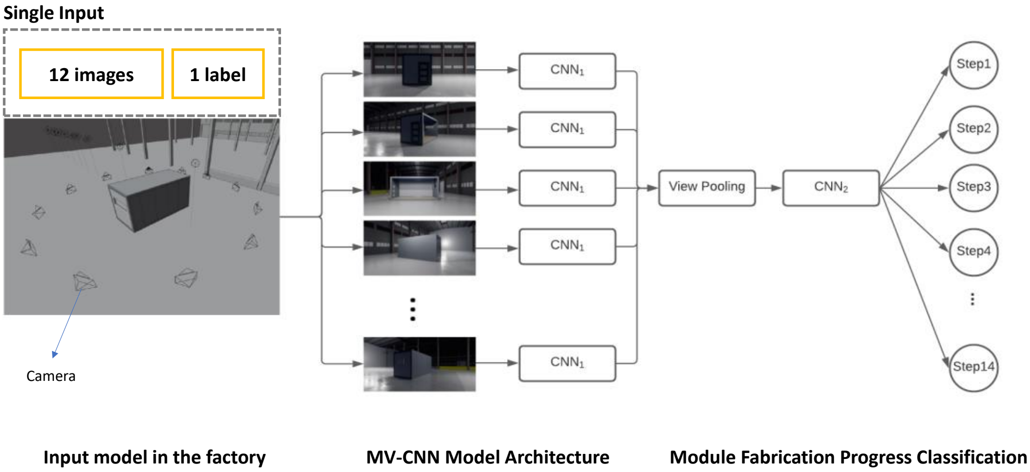


Figure 4. MV-CNN model architecture
12 images captured in VR per label (left); Processed into MV-CNN architecture (middle); Classification: each label is a step in the assembly sequence (right)

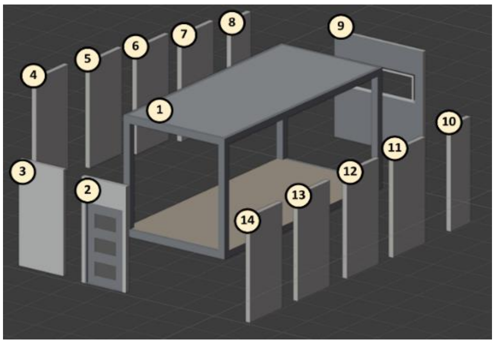


Figure 5. Module work pieces and numbered steps in the assembly sequence

Because the classification model identifies module progress within the given factory environment, we placed modules for each assembly step of the fabrication randomly on the assembly area in the virtual factory (Figure 6). These randomized scenes are providing different background visual representation of input image data (e.g., lighting, background objects, etc.) for models to learn under different contexts. In addition, the materials (texture, color) of these varying background objects (e.g., factory walls, floor, ceiling) have been configured to bring more variance to the context.

We generated 84,000 images which are 7,000 set of inputs (where each set has 12 images captured from different views per input). For each step of the assembly, we had 500 inputs, resulting in 500*12=6,000 images to use in the training. Therefore, each label (module assembly step) has 500 inputs that were separated into 400 training sets and 100 training sets.

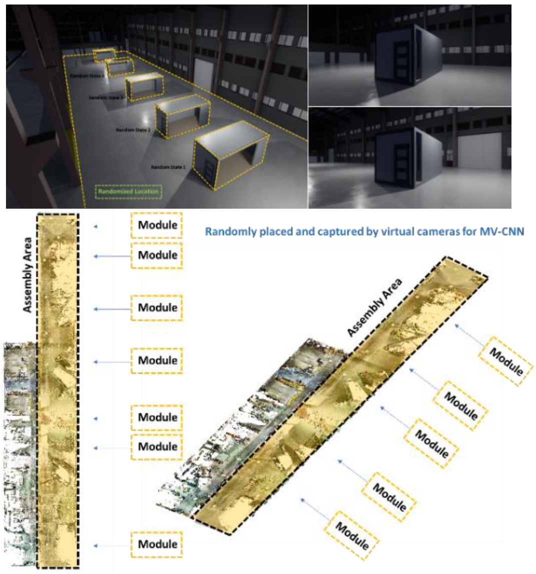


Figure 6. Randomized location of modules to capture images

4 Results

The accuracy, precision, recall, and f1-score of model result are 0.97. The testing results are shown in Table 1 as a confusion matrix. The prediction results of module assembly step 9 and step 13 are relatively lower than the other predicted labels (0.82 and 0.84, respectively). As the wrong predictions occurred later assembly steps, it is because majority viewpoints images are duplicated as completed parts are occluding the other parts (e.g., viewpoints image from 3 sides are same). Even though the background environments change, the model shows accurate results, which provides a strong evidence for suitability of using virtual environments for dataset generation.

Table 1. Confusion matrix for module assembly sequence classifier

	Predicted steps in the assembly sequence													
Act-ual	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	100													
2		100												
3			100											
4				100										
5					100									
6						100								
7							100							
8							2	96		2				
9									82	17			1	
10										97	1		2	
11										1	98	1		
12												100		
13												1	84	15
14														100

Since transfer learning is used, the training requires less computational resources and data. The accurate model with less cost through transfer learning shows that, the model can be tuned faster to adjust changes in volumetric units (or panels) to be assembled. This classifier can be used as a baseline to retrain classifiers for subsequent assembly sequences.

5 Conclusion

In this study, we evaluated the viability of training AI models for classifying module assembly sequences using the datasets generated in virtual environments within the modular construction context. Given the test dataset with possible variances, the model shows accurate results and provides a clear point of departure for utilization of synthetic datasets for training models. This is an essential step towards robotic assistance where robots intelligently assist human workers to bring the next required workpiece in the assembly sequence by understanding the step at which the assembly is at any point in time.

This paper provides the initial findings of an ongoing study and reports the following limitations. Since the training and test datasets were fully generated in virtual reality, verification in real-world settings is needed. We will evaluate the performance of this approach to generate and utilize datasets reflecting complex assembly lines (with various geometric representations and component types). We aim to utilize this approach for complicated module assembly sequences, where occlusions are more apparent and geometries are unconventional (e.g., spherical or cylindrical shaped work pieces).

References

[1] Lawson, R. M., Ogden, R. G., & Bergin, R. Application of modular construction in high-rise buildings. *Journal of architectural engineering*, 18(2), 148-154, 2012.

[2] Bock, T. The future of construction automation: Technological disruption and the upcoming ubiquity of robotics. *Automation in Construction*, 59, 113-121, 2015.

[3] Andrew Ng. “Issue 84”. Retrieved from https://www.deeplearning.ai/the-batch/issue-84/, access date: March 24, 2021

[4] Armeni, I., He, Z. Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J., & Savarese, S. 3d scene graph: A structure for unified semantics, 3d space, and camera. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5664-5673, 2019.

[5] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gorden, D., Zhu, Y., Gupta, A., & Farhadi, A. Ai2-thor: An interactive 3d environment for visual ai. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., & Zhang, Y. Matterport3d: Learning from rgb-d data in indoor environments. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] Xia, F., Zamir, A. R., He, Z., Sax, A., Malik, J., & Savarese, S. Gibson env: Real-world perception for embodied agents. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9068-9079, 2018.

[8] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, Y., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., & Newcombe, R. The Replica dataset: A digital replica of indoor spaces. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[9] Koo, B., Jung, R., & Yu, Y. Automatic classification of wall and door BIM element subtypes using 3D geometric deep neural networks. *Advanced Engineering Informatics*, 47, 101200, 2021.

[10] Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. *In Proceedings of the IEEE international conference on computer vision*, 945-953, 2015.

[11] Kim, H., & Mun, D. Deep-learning-based classification and retrieval of components of a process plant from segmented point clouds. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Park, K., & Ergan, S. Towards Intelligent Agents to Detect Work Pieces and Processes in Modular Construction: An Approach to Generate Synthetic Training Data. *ASCE CI and CRC Joint Conference 2022* (Submitted).

[13] Hamdi, A., Giancola, S., & Ghanem, B. MVTN: Multi-View Transformation Network for 3D Shape Recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[14] You, H., Feng, Y., Ji, R., & Gao, Y. Pvnet: A joint convolutional network of point cloud and multi-view for 3d shape recognition. *In Proceedings of the 26th ACM international conference on Multimedia*, 1310-1318, 2018.

[15] Qi, C. R., Su, H., Mo, K., & Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 652-660, 2017.

[16] Rashid, K. M., & Louis, J. Activity identification in modular construction using audio signals and machine learning. *Automation in Construction*, 119, 103361, 2020.

[17] Zheng, Z., Zhang, Z., & Pan, W. Virtual prototyping-and transfer learning-enabled module detection for modular integrated construction. *Automation in Construction*, 120, 103387, 2020.

[18] Zhou, Y., Huang, J., Dai, X., Liu, S., Luo, L., Chen, Z., & Ma, Y. HoliCity: A city-scale data platform for learning holistic 3D structures. *In Proceedings of the*

IEEE conference on computer vision and pattern recognition, 2020.

[19] Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., & Funkhouser, T. Semantic scene completion from a single depth image. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1746-1754, 2017.

[20] Fu, H., Cai, B., Gao, L., Zhang, L., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., & Zhang, H. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[21] Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., & Zhou, Z. Structured3d: A large photo-realistic dataset for structured 3d modeling. In Computer Vision–ECCV 2020: *16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16 (pp. 519-535). Springer International Publishing*, 2020.

[22] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828-5839, 2017.

[23] Yang, K., Ahn, C. R., & Kim, H. Deep learning-based classification of work-related physical load levels in construction. *Advanced Engineering Informatics*, 45, 101104, 2020.

[24] Rashid, K. M., & Louis, J. Times-series data augmentation and deep learning for construction equipment activity recognition. *Advanced Engineering Informatics*, 42, 100944, 2019.

[25] Kim, J., Song, J., & Lee, J. K. Recognizing and classifying unknown object in BIM using 2D CNN. *In International Conference on Computer-Aided Architectural Design Futures*, 47-57. Springer, Singapore, 2019.

[26] National Institute of Building Science (NIBS), Labor productivity index for US construction industry and all non-farm industries from 1964 through 2003, 2007.

[27] McKinsey & Company. The construction productivity imperative. Retrieved from https://www.mckinsey.com/business-functions/operations/our-insights/the-construction-productivity-imperative, access date: Jul 24, 2021